

FAST LANDMARK SUBSPACE CLUSTERING

XU WANG AND GILAD LERMAN
 DEPT. OF MATHEMATICS, UNIVERSITY OF MINNESOTA,
 MINNEAPOLIS, MN 55455, USA
 {WANG1591, LERMAN}@UMN.EDU

ABSTRACT. Kernel methods obtain superb performance in terms of accuracy for various machine learning tasks since they can effectively extract nonlinear relations. However, their time complexity can be rather large especially for clustering tasks. In this paper we define a general class of kernels that can be easily approximated by randomization. These kernels appear in various applications, in particular, traditional spectral clustering, landmark-based spectral clustering and landmark-based subspace clustering. We show that for n data points from K clusters with D landmarks, the randomization procedure results in an algorithm of complexity $O(KnD)$. Furthermore, we bound the error between the original clustering scheme and its randomization. To illustrate the power of this framework, we propose a new fast landmark subspace (FLS) clustering algorithm. Experiments over synthetic and real datasets demonstrate the superior performance of FLS in accelerating subspace clustering with marginal sacrifice of accuracy.

1. INTRODUCTION

Kernel-based learning algorithms have been highly successful for various machine learning tasks. Such algorithms typically contain two steps, first nonlinearly mapping the input data \mathcal{X} into a high-dimensional feature space \mathcal{F} and then applying linear learning algorithms on \mathcal{F} .

When designing kernel methods, it is of primary importance to make it applicable for large datasets. In other words, kernel methods need to scale linearly w.r.t. the number of data points. Unfortunately, most kernel methods require computation related to the kernel matrix. It scales poorly since the number of operations of merely obtaining the full kernel matrix is $O(n^2)$. Furthermore, many clustering algorithms require an eigen-decomposition of the kernel matrix.

Motivated by this scalability problem, we define a general class of kernels and study the properties of fast randomization approximations. This is a flexible framework that allows applications in different scenarios. In the setting of subspace clustering, we propose a landmark subspace clustering algorithm and show that it is fast with marginal sacrifice in accuracy.

1.1. Related Works. For building scalable kernel classification machines, Rahimi and Recht [18] proposed the approximation of kernels by Fourier random features and justified it by the Bochner's theorem of harmonic analysis. This technique was further analyzed and refined in other classification situations [13, 19, 10]. It has recently been shown to have comparable performance with deep learning [15] in both scalability and accuracy. However, Hamid et al. [9] observed that such random features for polynomial kernels contain redundant information, leading to

rank-deficient matrices. Although they are able to rectify it by dimension reduction, their work indicates that applying random features as suggested directly by the Bochner’s theorem may not be efficient.

In the territory of fast spectral clustering, the Nyström method in [14, 11] provides a low rank approximation of the kernel matrix by sampling over its columns. The error analysis of this approximation by special sampling schemes can be found in [2, 4, 12]. Different from the Nyström method, the landmark-based methods [1, 22, 7] first generate representative landmarks to summarize the underlying geometry of a dataset and then create new columns from these landmarks instead of sampling the original columns. Comparing with generating random Fourier features or subsampling the original columns, the landmark-based methods need less columns and thus are more efficient for spectral clustering.

1.2. Contributions of This Paper. First of all, we propose a randomization procedure that applies to a relatively large class of kernels, in particular, larger than classes treated by previous works (see Section 1.1). Second of all, we develop a kernel-based algorithm for fast approximated subspace clustering for a dataset of n points with K clusters. Given an approximation by D landmarks, the algorithm requires $O(KnD)$ running time. Third of all, in the context of kernel approximations via Fourier random features and landmarks, this is the first work that provides a bound of the L_2 -error between the original and approximated eigenvectors used in the clustering procedure. This bound is independent of n . We remark that Rahimi and Recht [18] only provided a bound of the difference between the corresponding entries of the original and approximated kernels.

1.3. Organization of This Paper. Section 2 defines a set of kernels that fit our approximation scheme and describes a randomized approximation procedure for such kernels. Section 3 introduces the fast landmark subspace (FLS) clustering algorithm. Section 4 shows that good approximation errors for both the kernel matrix and the eigenvectors of the normalized kernel matrix can be obtained by changing the number of landmarks with no dependence on the number of points. This analysis applies not only to FLS but also to other clustering algorithms associated with the different kernels introduced in Section 2. In Section 5, we compare the proposed FLS with state-of-the-art fast subspace clustering algorithms over various datasets.

2. KERNELS

2.1. Construction of Kernels. In this section we introduce a class of kernels. Let (X, μ_X) and (Y, μ_Y) be two measurable spaces and f be a bounded continuous map:

$$f: X \times Y \rightarrow \mathbb{R},$$

which is L_2 -integrable w.r.t. to the product measure $\mu_X \times \mu_Y$. This map induces an embedding

$$(1) \quad \phi: x \in X \mapsto f(x, \cdot) \in L^2(Y, d\mu_Y).$$

Moreover, if we define $k(x_1, x_2)$ on $X \times X$ as follows:

$$(2) \quad k(x_1, x_2) = \int_{y \in Y} f(x_1, y) f(x_2, y) d\mu_Y(y),$$

then the Fubini's theorem and [6, Page 6] imply that $k(x_1, x_2)$ is a Mercer's kernel. Indeed, for any $u(x) \in L^2(X, d\mu_X)$, the Mercer's condition is satisfied:

$$\begin{aligned} \int_{x_1, x_2 \in X} u(x_1)k(x_1, x_2)u(x_2)d\mu_X(x_1)d\mu_X(x_2) \\ = \int_{y \in Y} \left(\int_{x \in X} f(x, y)u(x)d\mu_X(x) \right)^2 d\mu_Y(y) \geq 0. \end{aligned}$$

We now show that different types of kernels can be constructed in this way.

Example I. The first example shows that different kernels can be obtained by varying the measure μ_Y . Let $X = \mathbb{R}^d$ and μ_X be the Lebesgue measure. Let $Y = \mathbb{R}^d \times [0, 2\pi]$ and μ_Y be the product of the Gaussian measure $N(\mathbf{0}, 1/\sigma^2 \mathbf{I})$ and the uniform measure $U([0, 2\pi])$. If we define

$$(3) \quad f(\mathbf{x}, (\mathbf{w}, t)) = \sqrt{2} \cos(\mathbf{w}^T \mathbf{x} + t),$$

then the corresponding kernel $k(\mathbf{x}_1, \mathbf{x}_2)$ defined in (2) is the usual Gaussian kernel with variance σ^2 . This formulation is a special case of the following Bochner's theorem (see also [18]) from harmonic analysis.

Theorem 2.1 (Bochner [20]). *A continuous shift-invariant kernel $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y})$ on \mathbb{R}^d is positive-definite if and only if $k(\delta)$ is the Fourier transform of a non-negative measure.*

The implication of Theorem 2.1 is that every positive-definite shift-invariant kernel $k(\mathbf{x}_1, \mathbf{x}_2)$ on \mathbb{R}^d can be written as the expectation (integral)

$$\mathbb{E}_{\mu \times U([0, 2\pi])}(f(\mathbf{x}_1, (\mathbf{w}, t))f(\mathbf{x}_2, (\mathbf{w}, t))),$$

where f is as in (3) for some non-negative measure μ . In other words, the set of kernels defined in (2) by choosing different measures μ_Y encompass all positive-definite shift-invariant kernels on \mathbb{R}^d .

Example II. The second example shows that this new formulation of kernels provides a theoretical foundation for the landmark-based methods [1, 22, 7]. Indeed, let X, Y be \mathbb{R}^d , μ_X be the Lebesgue measure and μ_Y be the probability measure from which the observed dataset is sampled. If we define

$$(4) \quad f(\mathbf{x}, \mathbf{y}) = (2\pi\sigma^2)^{-k/2} e^{-\|\mathbf{x} - \mathbf{y}\|_2^2 / (2\sigma^2)},$$

then the corresponding kernel $k(\mathbf{x}_1, \mathbf{x}_2)$ on X defined in (2) is the kernel implicitly used in landmark-based methods. Thus the theoretical analysis in Section 4 applies to these methods. We note that if one picks f as above and μ_Y be the Lebesgue measure, one obtains the Gaussian kernel on X .

2.2. Alternative Interpretation. The basic interpretation of this construction of kernels is that $f(x, \cdot)$ can be taken as an infinite dimensional representation of x in $L^2(Y, d\mu_Y)$ and $k(x_1, x_2)$ be the corresponding inner product. Alternatively speaking, each point $y \in Y$ provides a similarity score $f(x_1, y)f(x_2, y)$ for all pairs $(x_1, x_2) \in X \times X$. The kernel is an average of all such scores provided by the measurable space Y . This alternative interpretation provides an insight on the choice of μ_Y .

In Example II of Section 2, we mentioned that the integration of (4) over the underlying distribution of a given dataset leads to landmark-based kernels and the integration over the Lebesgue measure leads to the usual Gaussian kernels. It is

evident that landmark-based kernels are more efficient since they use the average scores of landmarks, which summarize the underlying geometry and thus provide more relevant similarity scores than points randomly chosen from the whole \mathbb{R}^d .

2.3. Randomized Approximation. In this section we present a randomized mapping ψ to approximate ϕ . We explicitly embed each $x \in X$ into an Euclidean space \mathbb{R}^D using a randomized mapping: $\psi : x \rightarrow \psi(x) \in \mathbb{R}^D$ where the inner product in \mathbb{R}^D approximates the kernel function. That is,

$$k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle_{\mu_Y} \approx \langle \psi(x_1), \psi(x_2) \rangle.$$

Let $\psi_y(x) = f(x, y)$. Then $\mathbb{E}_{\mu_Y}[\psi_y(x_1)\psi_y(x_2)] = k(x_1, x_2)$ for all $x_1, x_2 \in \mathbb{R}^d$ by definition. We randomly draw D points $y_i \in Y$ according to the probability measure μ_Y and define a random map:

$$(5) \quad \psi : x \in X \rightarrow \frac{1}{\sqrt{D}}[\psi_{y_1}(x), \dots, \psi_{y_D}(x)]^T \in \mathbb{R}^D.$$

3. FAST SUBSPACE CLUSTERING

3.1. Subspace Kernels. Different kernels can be obtained by taking a suitable measurable space Y in (2). In this section we show how this idea can be applied to the subspace clustering task, leading to a fast subspace clustering algorithm. Let the measurable space X be the sphere \mathbb{S}^d (after normalizing the data) and the measurable space Y be the Grassmannian $G(d, l)$ of l -dimensional subspaces with a measure μ_Y . We define the function

$$f : (x, L) \in X \times Y \mapsto \exp(-\text{dist}(x, L)^2/\sigma^2) \in \mathbb{R},$$

and the subspace kernel

$$(6) \quad k(x_1, x_2) = \int_{L \in Y} f(x_1, L)f(x_2, L)d\mu_Y(L).$$

We explain why this can be a good kernel for spectral subspace clustering in Section 2.2 below and further illustrate its power in the numerical section.

3.2. Choice of μ_Y . Lemma 3.1 shows that the uniform measure on the Grassmannian $G(d, l)$ is not a good choice. Indeed, the resulting kernel only has distance information between points and thus is a shift-invariant kernel in Example I of Section 2, which is not capable to detect the linear structures.

Lemma 3.1. *If μ_Y is the uniform measure on $G(d, l)$, then $k(x_1, x_2)$ is a function $g(\text{dist}(x_1, x_2))$ depending only on the Euclidean distance between x_1 and x_2 .*

Section 2.2 provides an answer for how to pick μ_Y for subspace kernels (6) given a subspace clustering task. Indeed, we note that the true underlying subspaces provide more relevant scores for the clustering task. A true subspace assigns 1 to a pair (x_1, x_2) if both of them belong to this subspace and a small score if otherwise. Therefore, a good choice of μ_Y is a measure supported on a neighborhood of the underlying subspaces of a given dataset in $G(d, l)$.

Landmark Subspace. Empirically the true subspaces are unknown and μ_Y can not be explicitly specified. Therefore the best we can do is to generate a set of subspaces $\{L_i\}_{i=1}^D$ that are close to the true subspaces in order to have a good approximation $k(x_1, x_2) \approx \frac{1}{D} \sum_{i=1}^D f(x_1, L_i) f(x_2, L_i)$. This set of subspaces is generated as follows. We first pick D landmark points for the dataset by random sampling or K -means. For each landmark point, a local best fit flat is found by selecting an optimal neighborhood radius according to [24]. These local flats approximate the true subspaces and are called landmark subspaces.

3.3. The FLS Algorithm. With the kernel approximation in Section 2.3, we can easily define the approximated normalized kernel matrix $\hat{\mathbf{L}}_{n,D}$ for a given dataset $X = \{\mathbf{x}_i\}_{i=1}^n$. Let

$$\hat{\mathbf{W}} = \psi(X)^T \psi(X) = [\psi(\mathbf{x}_1) \cdots \psi(\mathbf{x}_n)]^T [\psi(\mathbf{x}_1) \cdots \psi(\mathbf{x}_n)]$$

and the degree matrix $\hat{\mathbf{D}}$ be a diagonal matrix with the diagonal entries $\hat{D}_{ii} = \langle \psi(\mathbf{x}_i), \sum_{j=1}^n \psi(\mathbf{x}_j) \rangle$. Then we define

$$\hat{\mathbf{L}}_{n,D} = \hat{\mathbf{D}}^{-1/2} \hat{\mathbf{W}} \hat{\mathbf{D}}^{-1/2}.$$

The clustering is based on the top K eigenvectors of $\hat{\mathbf{L}}_{n,D}$. At first glance, there seems no reason to consider $\hat{\mathbf{L}}_{n,D}$ since its construction requires $O(n^2 D)$ operations. The essential idea of our randomized procedure is to provide a random low-rank decomposition of the normalized kernel matrix. The task of obtaining the top K eigenvectors of $\hat{\mathbf{L}}_{n,D}$ is transformed to finding the top K singular vectors of $\psi(X) \hat{\mathbf{D}}^{-1/2}$ since they are the same by definition. This can be done in $O(KnD)$. We formulate the algorithm as follows:

Algorithm 1 Fast landmark subspace clustering (FLS)

Input: $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \in \mathbb{R}^d$: data, d : dimension of subspaces, D : number of landmark subspaces, K : number of clusters, σ : scaling parameter for kernel, T, S : default parameters for local best-fit flats.

Output: Index set $\{g_i\}_{i=1}^N$ for K partitions such that $x_i \in X$ belongs to group with label g_i .

Steps:

- Generate D landmark points \mathbf{y}_j for X by either random sampling or applying k -means (see [1]).
- Find the local best-fit flat L_j (e.g., the landmark subspace) for each landmark point (cf., Algorithm 2 in [24]).
- Compute $\psi : \mathbf{x}_i \rightarrow [f(\mathbf{x}_i, L_1), \dots, f(\mathbf{x}_i, L_D)]^T, \quad \forall 1 \leq i \leq n$.
- Compute $\hat{D}_{ii} = \langle \psi(\mathbf{x}_i), \sum_{j=1}^n \psi(\mathbf{x}_j) \rangle$.
- Find the top K singular vectors $\{\hat{\mathbf{v}}_i\}_{i=1}^K$ for $\psi(X) \hat{\mathbf{D}}^{-1/2}$.
- Normalize each row of $[\hat{\mathbf{v}}_2, \dots, \hat{\mathbf{v}}_K]$ and call K -means on the rows.

return A cluster label g_i for each point \mathbf{x}_i .

4. THEORETICAL ANALYSIS

The main goal of the theoretical analysis is to bound the approximation errors of the kernel matrix entries and eigenvectors for the normalized kernel matrix. We show that the errors depend only on D . In other words, the number of landmarks

D does not have to increase as the dataset size n increase in order to achieve a fixed level of approximation precision. This rigorously justifies that the cost $O(KDn)$ of our algorithm is indeed linear in n without a hidden factor of n in D .

4.1. Uniform Convergence. Hoeffding's inequality guarantees the pointwise exponentially fast convergence in probability of $\langle \psi(x_1), \psi(x_2) \rangle = \frac{1}{D} \sum_{k=1}^D \psi_{y_k}(x_1) \psi_{y_k}(x_2)$ to $k(x_1, x_2)$. That is,

$$\Pr[|\psi(x_1)^T \psi(x_2) - k(x_1, x_2)| \geq \epsilon] \leq 2 \exp(-D\epsilon^2/4).$$

Recall that f is the function defining the kernel $k(x_1, x_2)$. If f and its derivatives in x are bounded by a constant C_f , we have the following stronger result that asserts the uniform convergence over a compact subset of X . Theorem 4.1 directly generalizes a similar argument in Claim 1 of [18] for more general manifolds X and the new class of kernels. For the completeness, we include its proof in the appendix.

Theorem 4.1. *Let \mathcal{M} be a compact subset of X . Then for the mapping ψ defined in (5),*

$$\Pr \left[\sup_{\mathbf{x}, \mathbf{y} \in \mathcal{M}} |\psi(\mathbf{x})^T \psi(\mathbf{y}) - k(\mathbf{x}, \mathbf{y})| \geq \epsilon \right] \leq C_{d,f} \exp(-D\epsilon^2/(16d+8))/\epsilon^2,$$

where $C_{d,f}$ is a constant depending on d and f and the covering number of \mathcal{M} . Furthermore, $\sup_{\mathbf{x}, \mathbf{y} \in \mathcal{M}} |\psi(\mathbf{x})^T \psi(\mathbf{y}) - k(\mathbf{x}, \mathbf{y})| \leq \epsilon$ with any constant probability when $D = \Omega(\frac{1}{\epsilon^2} \log(\frac{1}{\epsilon^2}))$.

We denote the kernel matrix by \mathbf{W} . Given a dataset $X = \{\mathbf{x}_i\}_{i=1}^n$, we recall that the approximation is given by

$$\hat{\mathbf{W}} = \psi(X)^T \psi(X) = [\psi(\mathbf{x}_1) \cdots \psi(\mathbf{x}_n)]^T [\psi(\mathbf{x}_1) \cdots \psi(\mathbf{x}_n)].$$

Theorem 4.1 states that the entrywise approximation quality of \mathbf{W} by $\hat{\mathbf{W}}$ is independent of n . This immediately indicates a low-rank approximation of \mathbf{W} as long as $D \ll n$.

4.2. Convergence of Eigenvectors. In many cases such as spectral clustering, the stability of eigenvectors is desired for a matrix approximation. For simplicity, we assume there are two clusters and consider the stability of the second largest eigenvector (the stability of the first eigenvector is trivial). The reasoning for top K eigenvectors is similar. Given a dataset $X = \{\mathbf{x}_i\}_{i=1}^n$, we recall that the normalized kernel matrix is defined as

$$\mathbf{L}_n = \mathbf{D}^{-1/2} \mathbf{W} \mathbf{D}^{-1/2},$$

where \mathbf{D} is diagonal with $D_{ii} = \sum_j W_{ij}$, and its approximation defined as

$$\hat{\mathbf{L}}_{n,D} = \hat{\mathbf{D}}^{-1/2} \hat{\mathbf{W}} \hat{\mathbf{D}}^{-1/2},$$

where $\hat{D}_{ii} = \langle \psi(\mathbf{x}_i), \sum_{j=1}^n \psi(\mathbf{x}_j) \rangle$. In this section, we show the convergence of the second largest eigenvector of $\hat{\mathbf{L}}_{n,D}$ to that of \mathbf{L}_n uniformly over n as $D \rightarrow \infty$. The spectral convergence related to $\hat{\mathbf{W}}$ and \mathbf{W} follows similarly. In the following analysis, we make the assumption below on the kernel $k(x_1, x_2)$:

Assumption 4.2. $k(x_1, x_2)$ is bounded from below and above by constants l, u ($0 < l < u$) on the domain from which the dataset is sampled.

We remark that this is also an essential assumption in proving the consistency of spectral clustering in [21] and Gaussian kernels satisfy this assumption over any compact set in \mathbb{R}^d .

Theorem 4.1 implies that $\hat{\mathbf{L}}_{n,D}$ entrywisely converges to \mathbf{L}_n as $D \rightarrow \infty$. Therefore, the second largest eigenvector $\hat{\mathbf{v}}_{1,n}$ of $\hat{\mathbf{L}}_{n,D}$ converges to the second largest eigenvector $\mathbf{v}_{1,n}$ of \mathbf{L}_n as $D \rightarrow \infty$ for each fixed n . In the following, we show that this convergence is uniform over n . That is, the approximation error depends only on D , even as $n \rightarrow \infty$.

The main idea is as follows. We first show that $\hat{\mathbf{L}}_{n,D}$ is a small perturbation of \mathbf{L}_n in operator L_2 -norm in Lemma 4.3. Then we show that the first nonzero eigenvalue of \mathbf{L}_n has strictly positive distance from the rest spectrum for all n in Lemma 4.4. This implies that a small perturbation of \mathbf{L}_n does not significantly change the corresponding eigenvector $\mathbf{v}_{1,n}$.

We now make the above arguments precise. To begin with, we assume the following probabilistic model. Suppose \mathbf{x}_i in X are i.i.d. sampled from the unit ball $B(\mathbf{0}, 1) \subset \mathbb{R}^d$ according to some probability measure P .

We denote $H_1 = (\mathbf{D}^{-1/2} - \hat{\mathbf{D}}^{-1/2})\mathbf{W}\mathbf{D}^{-1/2}$, $H_2 = \hat{\mathbf{D}}^{-1/2}(\mathbf{W} - \hat{\mathbf{W}})\mathbf{D}^{-1/2}$ and $H_3 = \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{W}}(\mathbf{D}^{-1/2} - \hat{\mathbf{D}}^{-1/2})$. It is easy to see that

$$\hat{\mathbf{L}}_{n,D} = \mathbf{L}_n - H_1 - H_2 - H_3.$$

The first task is to show that the L_2 -norms $\|H_i\|_2$ are small uniformly over n . We note that $\text{diam}(B(\mathbf{0}, 1)) = 2$. Theorem 4.1 implies that given any $\delta > 0$ and any probability $p > 0$ with $D := \frac{c_p}{\delta^2 l^2} \log \frac{1}{\delta l}$ where c_p is a constant depending only on p , the entrywise error $|W_{ij} - \hat{W}_{ij}| \leq \delta l$ with probability p . The first task is to show that $\|H_i\|_2, i = 1, 2, 3$ are small uniformly over n . This is formulated as follows and proved in the Appendix.

Lemma 4.3. *If $\delta > 0, p > 0$, $D := \frac{c_p}{\delta^2 l^2} \log \frac{1}{\delta l}$, where c_p is a constant depending only on p , then $\|H_1\|_2 \leq \frac{u}{2l}(1 - \delta)^{-3/2}\delta$, $\|H_2\|_2 \leq (1 - \delta)^{-1/2}\delta$ and $\|H_3\|_2 \leq \frac{1}{2(1-\delta)^2}(\frac{u}{l} + \delta)\delta$ with probability p , for all n .*

Lemma 4.3 shows that $\hat{\mathbf{L}}_{n,D}$ is a small perturbation of \mathbf{L}_n in L_2 -norm uniformly over n . Now we analyze further the relation between their spectrums. Let $\sigma(\mathbf{L}_n)$ and $\lambda_{1,n}$ be the spectrum and the second largest eigenvalue of \mathbf{L}_n . Lemma 4.4 below shows that $\lambda_{1,n}$ is sufficiently separated from the rest spectrum of \mathbf{L}_n .

Lemma 4.4. *Let $c > 0$ be a constant depending on the generating probability P of the dataset X and the kernel $k(x_1, x_2)$. For any constant probability p , there exists N_p such that*

$$\text{dist}(\lambda_{1,n}, \{0\} \cup \sigma(\mathbf{L}_n) \setminus \{\lambda_{1,n}\}) \geq c, \quad \forall n \geq N_p.$$

Lemma 4.4 together with the Davis-Kahan theorem [3] (see also [23, Theorem 1]) indicates that $\lambda_{1,n}$ and $\mathbf{v}_{1,n}$ are robust (not mixed with other eigenvectors) under a small perturbation of \mathbf{L}_n . Indeed, we have

$$\|\hat{\mathbf{v}}_{1,n} - \mathbf{v}_{1,n}\|_2 \leq \frac{\|H_1 + H_2 + H_3\|_2}{c} \leq C\delta,$$

for some constant $C > 0$ and $\delta < 1/2$. Thus Theorem 4.5 below follows.

Theorem 4.5. *For any $0 < \delta < 1/2$ and $0 < p \leq 1$, there are constants c_p and N_p such that if $D = \frac{c_p}{\delta^2 l^2} \log \frac{1}{\delta l}$ and $n > N_p$, then*

$$\|\hat{\mathbf{v}}_{1,n} - \mathbf{v}_{1,n}\|_2 \leq C\delta,$$

with probability at least p . Here C depends on the gap of top eigenvalues and the lower and upper bounds of the kernel $k(x_1, x_2)$ on the compact domain.

If there are K clusters, we need to consider first $K - 1$ nonzero eigenvectors $\{\mathbf{v}_{i,n}\}_{i=1}^{K-1}$ and their perturbations $\{\hat{\mathbf{v}}_{i,n}\}_{i=1}^{K-1}$. One can use similar arguments as above to prove the robustness of $\{\mathbf{v}_{i,n}\}_{i=1}^{K-1}$. We note that the approximation error of eigenvectors is controlled by the number of landmarks D .

5. EXPERIMENT

We compare FLS with the following algorithms: Local Best-fit Flats (LBF) [24], Sparse Subspace clustering (SSC) [5], Scalable Sparse Subspace Clustering (SSSC) [17]. LBF and SSSC are algorithms with the state-of-the-art performance in speed with reasonable high accuracy. Their comparison with other subspace clustering algorithms can be found in [24, 17]. Yet, there is no direct comparison between them. SSC is a popular but slow algorithm, which is included as a contrast for accuracy. We report the time cost in seconds and measure the accuracy of these algorithms by the ratio of correctly-clustered points with outliers excluded. That is,

$$\text{rate} = \frac{\# \text{ of correctly-clustered inliers}}{\# \text{ of total inliers}} \times 100\%.$$

When we fail to obtain a result due to the exceeding of memory limits, we report it as N/A.

5.1. Synthetic Data. FLS was tested and compared with other algorithms for artificial data with various subspace dimensions and levels of outliers. We use the notation $(d_1, \dots, d_K) \in \mathbb{R}^D$ to denote the model of K subspaces in \mathbb{R}^D of dimensions d_1, \dots, d_K . Given each model, we repeatedly generate 10 different sample sets according to the code in [16]. More precisely, for each subspace in the model, 250 points are first created by drawing uniformly at random from the unit disk of that subspace and then corrupted by Gaussian noise (e.g., from $N(\mathbf{0}, 0.05^2 \mathbf{I}_{D \times D})$ on \mathbb{R}^D). Then the whole sample set is further corrupted by adding 5% or 30% uniformly distributed outliers in a cube of side-length determined by the maximal distance of the former generated data to the origin. For each model, we repeat the experiment on 10 sample sets and the average accuracy (running time) are reported in Table 1 and 2 for outlier levels 5% and 30% respectively.

TABLE 1. Average accuracy (time (s)) for outlier level 5%.

	$(2, 2) \in \mathbb{R}^6$	$(4, 5, 6) \in \mathbb{R}^{10}$	$(5, 6, 7) \in \mathbb{R}^{20}$	$(3, 4, 5, 6, 7) \in \mathbb{R}^{80}$
FLS	0.99 (0.81)	0.98 (0.81)	1.00 (1.0)	1.00 (4.7)
LBF	0.99 (0.65)	0.97 (0.98)	1.00 (1.3)	0.98 (8.7)
SSC	0.92 (140)	0.55 (290)	0.98 (330)	0.95 (160)
SSSC	0.89 (2.6)	0.62 (4.2)	0.83 (6.8)	0.73 (6.4)

TABLE 2. Average accuracy (time (s)) for outlier level 30%.

	$(2, 2) \in \mathbb{R}^6$	$(4, 5, 6) \in \mathbb{R}^{10}$	$(5, 6, 7) \in \mathbb{R}^{20}$	$(3, 4, 5, 6, 7) \in \mathbb{R}^{80}$
FLS	0.99 (0.69)	0.98 (0.99)	1.00 (1.2)	0.99 (7.9)
LBF	0.99 (0.61)	0.98 (1.0)	1.00 (1.6)	0.98 (11)
SSC	0.91 (210)	0.59 (360)	0.84 (260)	0.73 (760)
SSSC	0.76 (2.6)	0.44 (4.1)	0.47 (6.2)	0.40 (6.9)

5.2. MNIST Dataset. We test the four algorithms on the MNIST dataset of handwritten digits. We work on the training set of 60000 28×28 images of the digits 0 through 9. We pick images of a combination of several digits and apply PCA to reduce the dimension to $D = 80$. We choose a fixed subspace dimension $d = 3$ and the correct number of clusters. The clustering rate is reported in Table 3. We note that $[1 - 9]$ stands for $[1, 2, 3, 4, 5, 6, 7, 8, 9]$ in the table.

TABLE 3. Average accuracy (time (s)) for MNIST dataset.

	$[1, 7]$	$[4, 9]$	$[2, 4, 8]$	$[3, 6, 8]$	$[2, 5, 7, 8, 9]$	$[1 - 9]$
FLS	0.99 (15)	0.93 (13)	0.97 (20)	0.95 (19)	0.81 (38)	0.54 (111)
LBF	0.65 (6)	0.57 (7)	0.44 (12)	0.57 (12)	0.36 (27)	0.25 (85)
SSC	0.76 (18993)	0.66 (14358)	0.94 (45683)	0.86 (47448)	N/A	N/A
SSSC	0.89 (33)	0.65 (27)	0.95 (34)	0.90 (37)	0.77 (70)	0.50 (152)

5.3. Other Datasets. We also tried the four algorithms on the Extended Yale B (ExtendYB) [8], the penDigits dataset from UCI database. The Extended Yale B dataset contains 2414 front-face images from 38 persons. In the experiment, we randomly select 8 persons, crop their images to 48×42 and further reduce the dimension by projecting to the first 80 principal components. As suggest in [24, page 12], we apply a crude whitening process to the projected data (removing the top two principal components) before applying FLS and LBF. We report SSC and SSSC without such whitening since they worked better when all 80 directions are included. The penDigits dataset contains 7494 data points of 16 features, each of which represents a digit from 0 to 9. The results are reported in Table 4.

TABLE 4. Average accuracy (time (s)).

	FLS	LBF	SSC	SSSC
ExtendYB	0.68 (1.79)	0.63 (4.36)	0.87 (259.66)	0.57 (15.02)
penDigits	0.68 (18.58)	0.53 (13.33)	N/A	0.51 (44.07)

6. CONCLUSION

In this paper we introduce an efficient framework to approximate a large class of kernels and consequently obtain faster clustering algorithms. In the context of clustering, this framework proposes a novel procedure for fast subspace clustering. Its complexity is $O(KnD)$ for n points with K clusters and D landmarks. Our theoretical analysis establishes such complexity for clustering algorithms associated with our framework. More importantly it bounds the L_2 -error between the original and approximated eigenvectors of kernel matrices, which applies to FLS, Fourier random features and other landmark-based methods.

REFERENCES

- [1] X. Chen and D. Cai. Large scale spectral clustering with landmark-based representation. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011*, 2011.
- [2] A. Choromanska, T. Jebara, H. Kim, M. Mohan, and C. Monteleoni. Fast spectral clustering via the nyström method. In *Algorithmic Learning Theory - 24th International Conference, ALT 2013, Singapore, October 6-9, 2013. Proceedings*, pages 367–381, 2013.
- [3] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7:1–46, 1970.
- [4] P. Drineas and M. W. Mahoney. On the nyström method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.
- [5] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2765–2781, 2013.
- [6] G. E. Fasshauer. Positive definite kernels: past, present and future. *Dolomites Research Notes on Approximation*, 4(Special Issue on Kernel Functions and Meshless Methods):21–63, 2011.
- [7] Z. Gan, C. Sha, and J. Niu. Fast spectral clustering with landmark-based subspace iteration. In *13th IEEE International Conference on Data Mining Workshops, ICDM Workshops, TX, USA, December 7-10, 2013*, pages 773–779, 2013.
- [8] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001.
- [9] R. Hamid, Y. Xiao, A. Gittens, and D. DeCoste. Compact random feature maps. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 19–27, 2014.
- [10] Purushottam Kar and Harish Karnick. Random feature maps for dot product kernels. In N. D. Lawrence and M. A. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, volume 22, pages 583–591, 2012.
- [11] S. Kumar, M. Mohri, and A. Talwalkar. On sampling-based approximate spectral decomposition. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, pages 553–560, 2009.
- [12] S. Kumar, M. Mohri, and A. Talwalkar. Sampling methods for the nyström method. *Journal of Machine Learning Research*, 13:981–1006, 2012.
- [13] Q. Le, T. Sarlos, and A. Smola. Fastfood - approximating kernel expansions in loglinear time. In *30th International Conference on Machine Learning (ICML)*, 2013.
- [14] M. Li, X. Lian, J. T. Kwok, and B. Lu. Time and space efficient spectral clustering via column sampling. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 2297–2304, 2011.
- [15] Z. Lu, A. May, K. Liu, A. B. Garakani, D. Guo, A. Bellet, L. Fan, M. Collins, B. Kingsbury, M. Picheny, and F. Sha. How to scale up kernel methods to be as good as deep neural nets. *CoRR*, abs/1411.4000, 2014.
- [16] Y. Ma, R. Vidal, and K. Huang. Website, 2014. perception.csl.illinois.edu/gpca.
- [17] X. Peng, L. Zhang, and Z. Yi. Scalable sparse subspace clustering. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 430–437, 2013.

- [18] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems*, pages 1177–1184, 2007.
- [19] A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems*, pages 1313–1320, 2008.
- [20] W. Rudin. *Fourier analysis on groups*. Wiley Classics Library. John Wiley & Sons, Inc., New York, 1990. Reprint of the 1962 original, A Wiley-Interscience Publication.
- [21] U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Ann. Statist.*, 36(2):555–586, 2008.
- [22] D. Yan, L. Huang, and M. I. Jordan. Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*, pages 907–916, 2009.
- [23] Y. Yu, T. Y. Wang, and R. J. Samworth. A useful variant of the davis-kahan theorem for statisticians. *ArXiv e-prints*, 2014.
- [24] T. Zhang, A. Szlam, Y. Wang, and G. Lerman. Hybrid linear modeling via local best-fit flats. *International Journal of Computer Vision*, 100(3):217–240, 2012.

APPENDIX A. PROOF OF LEMMA 3.1

Let (x_1, x_2) and (x'_1, x'_2) be two pairs on \mathbb{S}^d such that the angles $\theta(x_1, x_2) = \theta(x'_1, x'_2)$. Then there exists an orthogonal transformation \mathbf{R} such that $x'_1 = \mathbf{R}x_1$ and $x'_2 = \mathbf{R}x_2$. Thus,

$$\begin{aligned}
k(x'_1, x'_2) &= \int_{L \in Y} f(\mathbf{R}x_1, L) f(\mathbf{R}x_2, L) d\mu_Y(L) \\
&= \int_{L \in Y} f(x_1, \mathbf{R}^{-1}L) f(x_2, \mathbf{R}^{-1}L) d\mu_Y(L) \\
&= \int_{L \in Y} f(x_1, L) f(x_2, L) d\mu_Y(\mathbf{R}L) \\
&= \int_{L \in Y} f(x_1, L) f(x_2, L) d\mu_Y(L) = k(x_1, x_2).
\end{aligned}$$

We use the fact that μ_Y is uniform in the fourth equality. This equality means that $k(x_1, x_2) = g(\theta(x_1, x_2))$. We conclude the proof by noting that $\theta(x_1, x_2)$ and $\text{dist}(x_1, x_2)$ uniquely determine each other for any pair (x_1, x_2) on \mathbb{S}^d .

APPENDIX B. PROOF OF THEOREM 4.1

We denote $h(x_1, x_2) = \psi(x_1)\psi(x_2) - k(x_1, x_2)$. Let $L_f = \|\nabla h(x_1^*, x_2^*)\|$, where $(x_1^*, x_2^*) = \arg \max_{X \times X} \|\nabla h(x_1, x_2)\|$. Then

$$\mathbb{E}L_f^2 \leq \mathbb{E}\|\nabla(\psi(x_1)\psi(x_2))\|^2 \leq C_f^4.$$

The first inequality follows from the same argument in [18]. By Markov's inequality,

$$(7) \quad \mathbb{P}(L_f \geq \frac{\epsilon}{2r}) \leq \left(\frac{2rC_f^2}{\epsilon} \right)^2.$$

On the other hand, if \mathcal{M} is a d -dimensional manifold with diameter $\text{diam}(\mathcal{M})$, then $\mathcal{M} \times \mathcal{M}$ has dimension $2d$, diameter $\sqrt{2}\text{diam}(\mathcal{M})$ and r -net covering number $T = \left(\frac{12\sqrt{2}d\text{diam}(\mathcal{M})}{r} \right)^{4d}$. If $(x_1^{(i)}, x_2^{(i)})$ be the vertices of the r -net, then by

Hoeffding's inequality

$$(8) \quad \mathbb{P}[\cup_{i=1}^T \{|h(x_1^{(i)}, x_2^{(i)})| \geq \epsilon/2\}] \leq 2T \exp(-D\epsilon^2/8).$$

Equations (7), (8) imply that

$$\mathbb{P}[\sup_{\mathcal{M} \times \mathcal{M}} |h(x_1, x_2)| \leq \epsilon] \geq 1 - 2T \exp(-D\epsilon^2/8) - \left(\frac{2rC_f^2}{\epsilon}\right)^2.$$

If we pick r such that the second and third terms on the right-hand side are equal and simplifying the expression, then

$$\mathbb{P}[\sup_{\mathcal{M} \times \mathcal{M}} |h(x_1, x_2)| \leq \epsilon] \geq 1 - C_{d,f} \exp(-D\epsilon^2/(16d+8))/\epsilon^2,$$

where $C_{d,f}$ is a constant depending on d and the bound of f and its derivatives.

APPENDIX C. PROOF OF LEMMA 4.3

We first bound the L_2 -norms of each factor in H_i .

$$\begin{aligned} \|\mathbf{D}^{-1/2}\|_2 &= \max_i D_{ii}^{-1/2} = \left(\min_i D_{ii}\right)^{-1/2} \leq (nl)^{-1/2}, \\ \|\hat{\mathbf{D}}^{-1/2}\|_2 &= \left(\min_i \hat{D}_{ii}\right)^{-1/2} \leq (n(1-\delta)l)^{-1/2}, \\ \|\mathbf{W}\|_2 &\leq n \max_{i,j} |W_{ij}| \leq nu, \quad \|\hat{\mathbf{W}}\|_2 \leq nu + n\delta l \\ (9) \quad \|\mathbf{W} - \hat{\mathbf{W}}\|_2 &\leq n \max_{i,j} |W_{ij} - \hat{W}_{ij}| \leq n\delta l, \\ \|\mathbf{D}^{-1/2} - \hat{\mathbf{D}}^{-1/2}\|_2 &= \max_i |D_{ii}^{-1/2} - \hat{D}_{ii}^{-1/2}| \leq \frac{\max_i |D_{ii} - \hat{D}_{ii}|}{2(n(1-\delta)l)^{3/2}} \\ &\leq \frac{1}{2} \delta (1-\delta)^{-3/2} (nl)^{-1/2}. \end{aligned}$$

Here is a hint for the last inequality of (9). The function $f(x) = x^{-1/2}$ is a convex decreasing function over $x > 0$. Therefore, $|f(x) - f(y)| \leq f'(\min\{x, y\})|x - y|$. Then the last inequality in (9) follows from the fact that

$$\min_i \{\min\{D_{ii}, \hat{D}_{ii}\}\} \geq n(1-\delta)l.$$

The upper bounds of $\|H_i\|_2$ follow immediately from (9) and the inequality $\|AB\|_2 \leq \|A\|_2 \|B\|_2$ for any matrices A, B .

APPENDIX D. PROOF OF LEMMA 4.4

Let the linear operator T on $C(B(\mathbf{0}, 1))$ be defined as follows:

$$T(f)(\mathbf{x}) = \int_{\mathbf{y} \in B(\mathbf{0}, 1)} k(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) / \sqrt{d(\mathbf{x})d(\mathbf{y})} dP(\mathbf{y}),$$

where the degree function $d(\mathbf{x}) = \int_{\mathbf{y} \in B(\mathbf{0}, 1)} k(\mathbf{x}, \mathbf{y}) dP(\mathbf{y})$. Since T is a compact operator, its eigenvalue accumulation point is 0. It is easy to see that 1 is an eigenvalue of T with the eigenvector $\sqrt{d(\mathbf{y})}$. Luxburg et al. [21] showed that the eigenvalues and eigenvectors of \mathbf{L} converge to those of T with a convergence rate $O(\frac{1}{\sqrt{n}})$. Since \mathbf{L} has all its eigenvalues in $[0, 1]$, so is T . We summarize this in Lemma D.1

Lemma D.1. *T has all eigenvalues between $[0, 1]$. 1 is its eigenvalue. The only accumulation point of the eigenvalues is 0.*

Let $\sigma(T)$ and $\lambda_1 > 0$ be the spectrum and the second largest eigenvalue of T respectively. For simplicity, we assume λ_1 ($\lambda_{1,n}$) is a simple eigenvalue. Lemma D.1 implies the distance

$$\text{dist}(\lambda_1, \{0\} \cup \sigma(U) \setminus \{\lambda_1\}) = 2c > 0$$

for some $\delta > 0$. Theorem 15 in [21] implies that for any constant probability p , there exists N_p such that if $n > N_p$, the

$$\text{dist}(\lambda_{1,n}, \{0\} \cup \sigma(\mathbf{L}_n) \setminus \{\lambda_{1,n}\}) \geq c$$

with probability at least p .